CS 726: Homework #4

Posted: 03/08/2020, due: 03/27/2020 by 5pm on Canvas

Please typeset or write your solutions neatly! If we cannot read it, we cannot grade it.

Note: You can use the results we have proved in class – no need to prove them again.

Q 1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function and let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed, convex, and nonempty set. Recall the definition of the gradient mapping: $G_\eta(\mathbf{x}) = \eta(\mathbf{x} - P_{\mathcal{X}}(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})))$, where $P_{\mathcal{X}}(\cdot)$ denotes the Euclidean projection onto $\mathcal{X}$. Prove that $G_\eta(\cdot)$ it $(2\eta + L)$-Lipschitz continuous. [20pts]

Q 2. Consider a constrained minimization problem $\min_{\mathbf{u}\in\mathcal{X}} f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and $\mathcal{X} \subseteq \mathbb{R}^n$ is a hyper-rectangle: $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : x_i \in [a_i, b_i], \forall i \in \{1, 2, \ldots, n\}\}$, where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are vectors that satisfy $\forall i \in \{1, \ldots, n\} : a_i < b_i$.

(i) What is $P_{\mathcal{X}}(\mathbf{x})$? Write it in closed form. [10pts]

(ii) Define $\Delta_i(\mathbf{x}) := \nabla_i f(\mathbf{x})\mathbf{e}_i$, where $\mathbf{e}_i$ is the i^{th} standard basis vector (except for its i^{th} coordinate, all coordinates are equal to zero; the i^{th} coordinate equals one). Define:

$$T(\mathbf{x}, i) := \operatorname*{argmin}_{\mathbf{u}\in\mathcal{X}} \left\{ \langle \Delta_i(\mathbf{x}), \mathbf{u} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \right\}.$$

Express the gradient mapping $G_L(\mathbf{x})$ as a function of L, $\mathbf{x}$, and $T(\mathbf{x}, i)$, $i \in \{1, \ldots, n\}$. [10pts]

(iii) Consider the following method that starts from some initial point $\mathbf{x}_0 \in \mathcal{X}$ and updates its iterates $\mathbf{x}_k$ for $k \geq 0$ as:

$$i_k^* = \operatorname*{argmax}_{1\leq i\leq n} |(G_L(\mathbf{x}_k))_i|$$
$$\mathbf{x}_{k+1} = T(\mathbf{x}_k, i_k^*).$$

Prove the following sufficient descent property for this algorithm:

$$(\exists \alpha > 0)(\forall k \geq 0) : \quad f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha}{2}\|G_L(\mathbf{x}_k)\|_2^2.$$

What is the largest α for which this property holds? What can you say about convergence of this method if f is bounded below by some $\bar{f}$? [10pts]

Q 3. Consider a constrained minimization problem $\min_{\mathbf{u}\in\mathcal{X}} f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and M-Lipschitz, and $\mathcal{X} \subseteq \mathbb{R}^n$ is closed, convex, and nonempty. Assume that there exists $\mathbf{x}^* \in \mathcal{X}$ that minimizes f. Recall from the class that we use $\mathbf{g}_\mathbf{x}$ to denote an arbitrary subgradient of f at $\mathbf{x} \in \mathcal{X}$. Distances within the set $\mathcal{X}$ are measured using some norm $\|\cdot\|$ (e.g., an ℓ_p norm for $p \geq 1$). Norm that is dual to $\|\cdot\|$ is denoted by $\|\cdot\|_*$ and is defined by: $\|\mathbf{z}\|_* = \max_{\mathbf{u}:\|\mathbf{u}\|\leq 1} \langle \mathbf{z}, \mathbf{u}\rangle$. By the definition of a dual norm, you can immediately deduce that:

$$(\forall \mathbf{u}, \mathbf{z}) : \langle \mathbf{u}, \mathbf{z}\rangle \leq \|\mathbf{u}\|\|\mathbf{z}\|_*.$$

Note also that since f is M-Lipschitz w.r.t. the norm $\|\cdot\|$, we have, $\forall \mathbf{x}, \mathbf{y}$ and all $\mathbf{g}_\mathbf{x} \in \partial f(\mathbf{x})$:

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq M\|\mathbf{x} - \mathbf{y}\| \quad \text{and } \|\mathbf{g}_\mathbf{x}\|_* \leq M.$$

In this part, you will analyze the convergence of the Mirror Descent algorithm, defined by:

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{u}\in\mathcal{X}} \{\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{u} - \mathbf{x}_k\rangle + D_\psi(\mathbf{u}, \mathbf{x}_k)\}, \tag{MD}$$

where, as in previous assignments, $D_\psi(\mathbf{x}, \mathbf{y})$ denotes the Bregman divergence between $\mathbf{x}$ and $\mathbf{y}$ w.r.t. a continuously-differentiable function ψ.

Assume that ψ is 1-strongly convex w.r.t. $\|\cdot\|$, namely:

$$(\forall \mathbf{x}, \mathbf{y}): \quad \psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \langle \nabla\psi(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

(i) Using the first-order optimality in the definition of $\mathbf{x}_{k+1}$ and the three-point identity of Bregman divergences we proved in HW #2, show that:

$$(\forall \mathbf{u} \in \mathcal{X}): \quad a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{u}\rangle \leq D_\psi(\mathbf{x}^*, \mathbf{x}_k) - D_\psi(\mathbf{x}^*, \mathbf{x}_{k+1}) - D_\psi(\mathbf{x}_{k+1}, \mathbf{x}_k). \quad \text{[15pts]}$$

(ii) Use Part (i) to prove that:

$$a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \leq \frac{{a_k}^2 M^2}{2} + D_\psi(\mathbf{x}^*, \mathbf{x}_k) - D_\psi(\mathbf{x}^*, \mathbf{x}_{k+1}). \quad \text{[15pts]}$$

(iii) Use Part (ii) to prove that, $\forall k \geq 0$:

$$f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*) \leq \frac{D_\psi(\mathbf{x}^*, \mathbf{x}_0) + \frac{M^2}{2}\sum_{i=0}^k {a_i}^2}{A_k},$$

where $\mathbf{x}_k^{\text{out}} = \frac{\sum_{i=0}^k a_i \mathbf{x}_i}{A_k}$. [10pts]

(iv) Discuss how you would choose $\{a_i\}_{i=0}^k$ for (MD) to converge as fast as possible, and provide the convergence bound (a bound on $f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*)$) for your choice(s) of $\{a_i\}_{i=0}^k$. [10pts]

Q 1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function and let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed, convex, and nonempty set. Recall the definition of the gradient mapping: $G_\eta(\mathbf{x}) = \eta(\mathbf{x} - P_{\mathcal{X}}(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})))$, where $P_{\mathcal{X}}(\cdot)$ denotes the Euclidean projection onto $\mathcal{X}$. Prove that $G_\eta(\cdot)$ it $(2\eta + L)$-Lipschitz continuous. [20pts]

$\forall x, y \in \mathcal{X}$.

$$G_\eta(x) - G_\eta(y) = \eta\left[x - y - \left(P_{\mathcal{X}}\left(x - \tfrac{1}{\eta}\nabla f(x)\right) - P_{\mathcal{X}}\left(y - \tfrac{1}{\eta}\nabla f(y)\right)\right)\right]$$

$$\left\| G_\eta(x) - G_\eta(y) \right\| \le \eta\|x - y\| + \eta\left\| P_{\mathcal{X}}\left(x - \tfrac{1}{\eta}\nabla f(x)\right) - P_{\mathcal{X}}\left(y - \tfrac{1}{\eta}\nabla f(y)\right)\right\|$$

$$\le \eta\|x - y\| + \eta\left\| x - y + \tfrac{1}{\eta}\left(\nabla f(y) - \nabla f(x)\right)\right\|$$

$$\le 2\eta\|x - y\| + \|\nabla f(x) - \nabla f(y)\|$$

$$\le (2\eta + L)\|x - y\|$$

Q 2. Consider a constrained minimization problem $\min_{\mathbf{u}\in\mathcal{X}} f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and $\mathcal{X} \subseteq \mathbb{R}^n$ is a hyper-rectangle: $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : x_i \in [a_i, b_i], \forall i \in \{1, 2, \ldots, n\}\}$, where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are vectors that satisfy $\forall i \in \{1, \ldots, n\} : a_i < b_i$.

(i) What is $P_{\mathcal{X}}(\mathbf{x})$? Write it in closed form. $\min\left\{\max\{a_i, x_i\}, b_i\right\}$ [10pts]

(ii) Define $\Delta_i(\mathbf{x}) := \nabla_i f(\mathbf{x})\mathbf{e}_i$, where $\mathbf{e}_i$ is the i^{th} standard basis vector (except for its i^{th} coordinate, all coordinates are equal to zero; the i^{th} coordinate equals one). Define:

$$T(\mathbf{x}, i) := \operatorname*{argmin}_{\mathbf{u}\in\mathcal{X}} \left\{ \langle \Delta_i(\mathbf{x}), \mathbf{u} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \right\}.$$

$u - x = -\frac{\partial_i(x)}{L}$

Express the gradient mapping $G_L(\mathbf{x})$ as a function of L, $\mathbf{x}$, and $T(\mathbf{x}, i)$, $i \in \{1, \ldots, n\}$. [10pts]

(iii) Consider the following method that starts from some initial point $\mathbf{x}_0 \in \mathcal{X}$ and updates its iterates $\mathbf{x}_k$ for $k \ge 0$ as:

$$i_k^* = \operatorname*{argmax}_{1\le i\le n} |(G_L(\mathbf{x}_k))_i|$$
$$\mathbf{x}_{k+1} = T(\mathbf{x}_k, i_k^*).$$

Prove the following sufficient descent property for this algorithm:

$$(\exists \alpha > 0)(\forall k \ge 0): \quad f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \frac{\alpha}{2}\|G_L(\mathbf{x}_k)\|_2^2.$$

What is the largest α for which this property holds? What can you say about convergence of this method if f is bounded below by some $\bar{f}$? [10pts]

(i) $P_{\mathcal{X}}(x) = \min\{\max\{a_i, x_i\}, b_i\}$

(ii) Let $\bar{x} = \operatorname{argmin}_{u\in\mathcal{X}}\left\{\langle\nabla f(x), u - x\rangle + \frac{L}{2}\|u - x\|_2^2\right\} = P_{\mathcal{X}}\left(x - \frac{1}{L}\nabla f(x)\right) = x - \frac{1}{L}G_L(x)$

$$G_L(x) = L\left(x - P_{\mathcal{X}}\left(x - \tfrac{1}{L}\nabla f(x)\right)\right)$$

We have:

$$T(x,i) = \operatorname{argmin}_{u\in\mathcal{X}} \left\{ \langle \Delta_i(x), u-x\rangle + \frac{L}{2}\|u-x\|_2^2 \right\} = P_{\mathcal{X}}\left(x - \frac{\Delta_i(x)}{L}\right) = x - \frac{1}{L} G_L(x)\cdot e_i$$

define $G_L^i(x) = G_L(x)\cdot e_i$

$$= L\left(x - T(x,i)\right)$$

$$G_L(x) = Lx - L\cdot \begin{bmatrix} T(x,1) \\ \vdots \\ T(x,n) \end{bmatrix}$$

(iii) $x_{k+1} = x_k - \frac{1}{L} G_L^{i_k^*}(x_k)$ where $i_k^* = \operatorname{argmax}_{0\le i\le n} \left|[G_L(x)]_i\right|$

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2$$

$$\le f(x_k) - \frac{1}{L}\langle \nabla f(x), G_L^{i_k^*}(x_k)\rangle + \frac{1}{2L}\|G_L^{i_k^*}(x_k)\|_2^2$$

$$= f(x_k) - \frac{1}{L}\langle \Delta_{i_k^*}(x), G_L^{i_k^*}(x_k)\rangle + \frac{1}{2L}\|G_L^{i_k^*}(x_k)\|_2^2$$

$$= f(x_k) - \frac{1}{L}\langle \Delta_{i_k^*}(x) - G_L^{i_k^*}(x_k), G_L^{i_k^*}(x_k)\rangle - \frac{1}{2L}\|G_L^{i_k^*}(x_k)\|_2^2$$

Since: $\langle \Delta_{i_k^*}(x) - L(x - T(x,i_k^*)), L(x - T(x,i_k^*))\rangle$

$$= L^2\langle -\left(x - \tfrac{1}{L}\Delta_{i_k^*}(x)\right) + T(x,i_k^*), x - T(x,i_k^*)\rangle = I$$

Consider: $T(x,i_k^*) = P_{\mathcal{X}}(x - \frac{1}{L}\Delta_{i_k^*})$

According to the property of Euclidean Projection:

$$\langle x - P_{\mathcal{X}}(x), u - P_{\mathcal{X}}(x)\rangle \le 0 \qquad \forall u\in\mathcal{X}$$

Then: $I = L^2\langle P_{\mathcal{X}}(x - \frac{1}{L}\Delta_{i_k^*}) - (x - \frac{1}{L}\Delta_{i_k^*}), x - P_{\mathcal{X}}(x - \frac{1}{L}\Delta_{i_k^*})\rangle \ge 0$

We have: $f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|G_L(x_k)\cdot e_{i_k^*}\|_2^2$

$$\le f(x_k) - \frac{1}{2nL}\|G_L(x_k)\|_2^2$$

$\alpha = \frac{1}{nL}$

If $f(x) \geq \bar{f} \quad \forall x$

$$f(x_{k+1}) - f(x_0) \leq \frac{-1}{2nL} \sum_{i=0}^{k} \| G_L(x_i) \|_2^2$$

$$\min_{0 \leq i \leq k} \| G_L(x_i) \|_2^2 \leq \frac{2nL}{k+1} \left(f(x_0) - f(x_{k+1}) \right) \leq \frac{2nL}{k+1} \left(f(x_0) - \bar{f} \right)$$

Q 3. Consider a constrained minimization problem $\min_{\mathbf{u} \in \mathcal{X}} f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and M-Lipschitz, and $\mathcal{X} \subseteq \mathbb{R}^n$ is closed, convex, and nonempty. Assume that there exists $\mathbf{x}^* \in \mathcal{X}$ that minimizes f. Recall from the class that we use $\mathbf{g_x}$ to denote an arbitrary subgradient of f at $\mathbf{x} \in \mathcal{X}$. Distances within the set $\mathcal{X}$ are measured using some norm $\| \cdot \|$ (e.g., an ℓ_p norm for $p \geq 1$). Norm that is dual to $\| \cdot \|$ is denoted by $\| \cdot \|_*$ and is defined by: $\|\mathbf{z}\|_* = \max_{\mathbf{u}:\|\mathbf{u}\| \leq 1} \langle \mathbf{z}, \mathbf{u} \rangle$. By the definition of a dual norm, you can immediately deduce that:

$$(\forall \mathbf{u}, \mathbf{z}) : \langle \mathbf{u}, \mathbf{z} \rangle \leq \|\mathbf{u}\| \|\mathbf{z}\|_*.$$

Note also that since f is M-Lipschitz w.r.t. the norm $\| \cdot \|$, we have, $\forall \mathbf{x}, \mathbf{y}$ and all $\mathbf{g_x} \in \partial f(\mathbf{x})$:

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq M \|\mathbf{x} - \mathbf{y}\| \quad \text{and } \|\mathbf{g_x}\|_* \leq M.$$

In this part, you will analyze the convergence of the Mirror Descent algorithm, defined by:

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{u} \in \mathcal{X}} \{ \langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{u} - \mathbf{x}_k \rangle + D_\psi(\mathbf{u}, \mathbf{x}_k) \}, \tag{MD}$$

$x_{k+1} = \operatorname{argmin}_{u \in \mathcal{X}} \{ \alpha_k \langle g_{x_k}, u - x_k \rangle + D_\psi(u, x_k) \}$

$\langle -\alpha_k g_{x_k} - \nabla D_\psi(x_{k+1}, x_k), u - x_{k+1} \rangle \leq 0 \quad \forall u$

$\langle \alpha_k g_{x_k} + \nabla_{x_{k+1}} D_\psi(x_{k+1}, x_k), x_{k+1} - u \rangle \leq 0 \quad \forall u$

where, as in previous assignments, $D_\psi(\mathbf{x}, \mathbf{y})$ denotes the Bregman divergence between $\mathbf{x}$ and $\mathbf{y}$ w.r.t. a continuously-differentiable function ψ.

$\langle \alpha_k g_{x_k} + \nabla \psi(x_{k+1}) - \nabla \psi(x_k), x_{k+1} - u \rangle \geq 0 \quad \forall u$

Assume that ψ is 1-strongly convex w.r.t. $\| \cdot \|$, namely:

$$(\forall \mathbf{x}, \mathbf{y}) : \quad \psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \langle \nabla \psi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

$\psi(y) - \psi(x) - \langle \nabla \psi(x), y - x \rangle = D_\psi(y, x) \geq \frac{1}{2} \|y - x\|_2^2$

(i) Using the first-order optimality in the definition of $\mathbf{x}_{k+1}$ and the three-point identity of Bregman divergences we proved in HW #2, show that:

$\langle \nabla \psi(x_{k+1}) - \nabla \psi(x_k), x^* - x_{k+1} \rangle$

$$(\forall \mathbf{u} \in \mathcal{X}) : \quad a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{u} \rangle \leq D_\psi(\mathbf{x}^*, \mathbf{x}_k) - D_\psi(\mathbf{x}^*, \mathbf{x}_{k+1}) - D_\psi(\mathbf{x}_{k+1}, \mathbf{x}_k). \quad \text{[15pts]}$$

(ii) Use Part (i) to prove that:

$$a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle \leq \frac{{a_k}^2 M^2}{2} + D_\psi(\mathbf{x}^*, \mathbf{x}_k) - D_\psi(\mathbf{x}^*, \mathbf{x}_{k+1}). \quad \text{[15pts]}$$

(iii) Use Part (ii) to prove that, $\forall k \geq 0$:

$$f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*) \leq \frac{D_\psi(\mathbf{x}^*, \mathbf{x}_0) + \frac{M^2}{2} \sum_{i=0}^{k} {a_i}^2}{A_k},$$

where $\mathbf{x}_k^{\text{out}} = \frac{\sum_{i=0}^{k} a_i \mathbf{x}_i}{A_k}$. [10pts]

(iv) Discuss how you would choose $\{a_i\}_{i=0}^k$ for (MD) to converge as fast as possible, and provide the convergence bound (a bound on $f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*)$) for your choice(s) of $\{a_i\}_{i=0}^k$. [10pts]

(i) Since $x_{k+1} = \underset{u \in \mathcal{X}}{\operatorname{argmin}} \{ a_k \langle g_{x_k}, u - x_k \rangle + D_\psi(u, x_k) \}$

$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$

we have $\langle - a_k g_{x_k} - \nabla_v D_\psi(v, x_k)\big|_{v = x_{k+1}}, u - x_{k+1} \rangle \le 0 \quad \forall u \in \mathcal{X}$

$- \langle a_k g_{x_k} + \nabla \psi(x_{k+1}) - \nabla \psi(x_k), x_{k+1} - u \rangle \ge 0$

$a_k \langle g_{x_k}, x_{k+1} - u \rangle \le \langle \nabla \psi(x_{k+1}) - \nabla \psi(x_k), u - x_{k+1} \rangle \quad \forall u \in \mathcal{X}$ ①

According to three-point identity of Bregman-divergence:

$\langle \nabla \psi(x_{k+1}) - \nabla \psi(x_k), u - x_{k+1} \rangle = D_\psi(u, x_k) - D_\psi(x_{k+1}, x_k) - D_\psi(u, x_{k+1})$

where $u \in \mathcal{X}$ ②

Since $f(y) \ge f(x) + \langle g_x, y - x \rangle$. Let $y = x + \lambda p \quad \forall p$ there $\|p\| = 1$

$\langle g_x, p \rangle \le \dfrac{f(x + \lambda p) - f(x)}{\lambda} = \langle \nabla f(x), p \rangle$ if $\nabla f(x)$ exists

$\langle g_x - \nabla f(x), p \rangle \le 0$ for $\forall p$.

then if $\nabla f(x)$ exists, $g_x = \nabla f(x)$ ③

Then combine ①. ②. ③.

$a_k \langle \nabla f(x_k), x_{k+1} - u \rangle \le D_\psi(x^*, x_k) - D_\psi(x_{k+1}, x_k) - D_\psi(x^*, x_{k+1}) \quad \forall u \in \mathcal{X}$

(ii) According to (i). we have:

$a_k \langle \nabla f(x_k), x_{k+1} - x^* \rangle \le D_\psi(x^*, x_k) - D_\psi(x_{k+1}, x_k) - D_\psi(x^*, x_{k+1})$ ①

According to $\psi(x)$ is 1- strongly convex: $\forall y, x \in \mathcal{X}$

$f(y) - f(x) - \langle \nabla \psi(x), y - x \rangle \ge \frac{1}{2} \|y - x\|_v^2$

$D_\psi(y, x) \ge \frac{1}{2} \|y - x\|_v^2$ ②

Then we have:

$a_k \langle \nabla f(x_k), x_k - x^* \rangle \le D_\psi(x^*, x_k) - D_\psi(x^*, x_{k+1}) - \frac{1}{2} \|x_{k+1} - x_k\|_v^2$

$+ a_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle$ ③

Claim: $-\frac{1}{2}\|x_{k+1}-x_k\|_v^2 + a_k\langle \nabla f(x_k), x_k - x_{k+1}\rangle \le \frac{a_k^2 M^2}{2}$

pf: $a_k\langle \nabla f(x_k), x_k - x_{k+1}\rangle = a_k\langle g_{x_k}, x_k - x_{k+1}\rangle \le a_k\|g_{x_k}\|\cdot\|x_k - x_{k+1}\|$

$\le a_k M\cdot\|x_k - x_{k+1}\|$

Let $p = \|x_{k+1}-x_k\|$ $\quad q = a_k M$

$-\frac{1}{2}\|x_{k+1}-x_k\|_v^2 + a_k\langle \nabla f(x_k), x_k - x_{k+1}\rangle - \frac{p^2}{2} + pq \le \frac{q^2}{2} = \frac{a_k^2 M^2}{2}$

then complete the proof of claim.

Then complete the proof of (ii). ▨

(iii) Consider optimality gap:

$U_k = \frac{1}{A_k}\sum_{i=0}^{k} a_i f(x_i) \ge f\left(\frac{1}{A_k}\sum_{i=0}^{k} a_i x_i\right) = f(x_k^{bmt})$

$A_k U_k - A_{k-1}U_{k-1} = a_k f(x_k)$

$f(x^*) \ge \frac{1}{A_k}\sum_{i=0}^{k} a_i\left(f(x_i) + \langle g_{x_i}, x^* - x_i\rangle\right) = L_k$

$A_k L_k - A_{k-1}L_{k-1} = a_k f(x_k) - a_k\langle g_{x_k}, x^* - x_k\rangle$

Then: $A_k G_k - A_{k-1}G_{k-1} = -a_k\langle g_{x_k}, x^* - x_k\rangle$

$\le \frac{a_k^2 M^2}{2} + D_\psi(x^*, x_k) - D_\psi(x^*, x_{k+1})$

$A_0 G_0 = a_0 f(x_0) - a_0 f(x_0) - a_0\langle g_{x_0}, x^* - x_0\rangle \le \frac{a_0^2 M^2}{2} D_\psi(x^*, x_0) - D_\psi(x^*, x_1)$

Thus:

$A_k G_k \le \frac{M^2}{2}\sum_{i=0}^{k} a_i^2 + D_\psi(x^*, x_0) - D_\psi(x^*, x_{k+1})$

$f(x_k^{bmt}) - f(x^*) \le G_k \le \dfrac{D_\psi(x^*, x_0) + \frac{M^2}{2}\sum_{i=0}^{k} a_i^2}{A_k}$

(iv): Choose $a_i = 2$

$$G_K \le \frac{D_\psi(x^*, x_0) + M^2 \alpha^2 (k+1)}{2\alpha(k+1)} = \frac{D_\psi(x^*, x_0)}{2\alpha(k+1)} + \frac{\alpha M^2}{2}$$

Choose α s.t. $\frac{D_\psi(x^*, x_0)}{2\alpha(k+1)} = \frac{\alpha M^2}{2} \Rightarrow \alpha = \sqrt{\frac{D_\psi(x^*, x_0)}{M^2(k+1)}}$

Then: $G_K \le \frac{\alpha M^2}{}$

$$= \sqrt{\frac{D_\psi(x^*, x_0)}{k+1}} \cdot M \le \varepsilon$$

$f(x_K^{avg}) - f(x^*) \le \varepsilon$ in

$$K = O\left(\frac{D_\psi(x^*, x_0)\, M^2}{\varepsilon^2}\right) \text{ iterations}$$